DOCUMENT RESUME

ED 161 377 HE 010 610

AUTHOR Baltes, Kenneth G.; Hendrix, Vernon L.
TITLE A Methodology for Data Structure Assessment in Higher Education Administration. AIR Forum Paper 1978.
INSTITUTION Minnesota Univ., Minneapolis.
PUB DATE May 78
NOTE 16p.; Paper presented at the annual Association for Institutional Research Forum (18th, Houston, Texas, May 21-25, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Cluster Analysis; *Data Analysis; *Data Bases; Data Collection; Data Processing; Educational Administration; Grouping Procedures; Higher Education; *Information Needs; *Management Information Systems; Methods; Models; *University Administration

ABSTRACT

Two recent developments in management information system technology and higher education administration have brought about the need for this study, designed to develop a methodology for revealing a relational model of the data base that administrators are operating from currently or would like to be able to operate from in the future. Administrations of higher education have been forced to rely more heavily on information systems to respond to the demands for accountability and allocations of limited resources. Information systems technology through the advent of data base management systems is able to be more responsive to administrative information needs, provided the relationships within the data required by administrators is known. The analysis, conducted at the University of Minnesota, consisted of testing several data grouping techniques including four hierarchical clustering methods, factor analysis, and observation of summary matrices on the data. Complete linkage and average linkage cluster analysis provided what appeared to be the most reliable groupings of the entities and were applied to the data. The methodology does reveal the relationships that respondents perceive to be in the data. The methodology as it was tested was effective as an aid to the data base designer in establishing a relational model of the data base. (Author/JMD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED161377

THE ASSOCIATION FOR INSTITUTIONAL RESEARCH

This paper is part of the ERIC Collection of 1978 Association for Institutional Research Forum Papers. All papers in the Collection were recommended for inclusion by the AIR Forum Publications Review Committee.

Robert H. Fenske
Arizona State University
(Editor, AIR Forum Publications)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

HE 010 610

A Methodology for Data Structure Assessment in Higher Education Administration

Kenneth G. Baltes
Assistant to the Vice President for Finance
Office of the Vice President for Finance
335 Morrill Hall
University of Minnesota
Minneapolis, Minnesota
(376-4628)

and

Vernon L. Hendrix
Professor, Department of Educational Administration
425 Health Services Building
University of Minnesota
St. Paul, Minnesota
(373-5568)

AIR Convention
Houston, 1978

Abstract

Two recent developments in management information system technology and higher education administration have brought about the need for this study. Administrations of higher education have been forced to rely more heavily on information systems to respond to the demands for accountability and allocation of limited resources. Information systems technology through the advent of database management systems is able to be more responsive to administrative information needs, provided the relationships within the data required by administrators is known. The problem is to develop a methodology for revealing a relational model of the database which administrators are operating from currently or would like to be able to operate from in the future.

The analysis consisted of testing several data grouping techniques including four hierarchical clustering methods, factor analysis and observation of summary matrices on the data. Complete linkage and average linkage cluster analysis provided what appeared to be the most reliable groupings of the entities. These two clustering techniques were applied to the data.

The methodology does reveal the relationships which respondents perceive to be in the data. The methodology may or may not have revealed top administrators' views of the data, depending on how well administrative staff members understand their administrator's views. The methodology as it was tested was effective as an aid to the database designer in establishing a relational model of the database.

A Methodology for Data Structure Assessment in Higher Education Administration

Introduction

Data base technology has brought with it a whole new field of study - the study of data. We now understand from the work of Codd, Martin, Date and others that data has an inherent structure which can be traversed, manipulated and modeled (Codd, 1970; Date, 1975; Martin, 1975). We also know that databases designed using the inherent data structure as a model will probably require less maintenance over time and be more responsive to current as well as future user needs.

However, the data processing industry has not yet developed an efficient relational database management system for large databases (by relational system here it is meant one which records all relationships in the data structure).

Problem

The problem then is to develop a method for determining which of the relationships in a given data structure will be most frequently used and therefore reflected in the efficiently designed data base. This paper disucsses a methodology for determining the relational needs of the top administration of a large university.

Objectives

The objectives of this study were to test the effectiveness of a variety of clustering techniques, with associated data collection procedures, for:

1. discovering the latent data element structures in the minds of higher education administrators based on their decision-making and information needs, and

2. communicating these structures to technical experts in the design of management information systems and data bases.

Success in these objectives would permit the optimization of relationships among data elements in the development of a new management information system or the restructuring of an existing system.

Procedure

The methodology is a multistep process which includes: Data Element collection; normalization to identify entities; entity grouping by administrative users; and analysis of the groupings. The methodology was tested at the University of Minnesota using major data files currently in use (for a more detailed discussion see Baltes, 1977.

Data Element Collection

The relative stability of a mature educational institution with well established conventional data systems led to the assumption that the data elements required in the database system will be, for the most part, elements which are already being collected through current data processing systems. Fifteen master files were identified by the systems analysts in the data processing department which contained most of the data elements used in current systems. Although there are some 400 files currently in use, those not included in this study are either a subset of one of the fifteen files selected; a different generation of one of the files selected; similar to one of the fifteen selected but in a different sequence; or a data file from a system not considered to be an integral part of

the University's administrative systems, such as the printing plant inventory file of the telephone billing file.

These fifteen files contained a total of approximately eleven hundred data items. The names and descriptions of the elements were stored through a computerized data dictionary for ease of access and manipulation. The data items were also described on a deck of cards which was used in the normalization process described in the next section.

Normalization

E.F. Codd defined the normalization technique as a method for grouping data items into a set of relations, producing a relational model. The normalization technique was used with the data items collected in the previous step in order to group the items into a manageable set of entities to be used in the survey described below.

The process of normalizing the eleven hundred items included the following steps:

(1) Data items whose function was primarily data processing related and contained no information for the user were eliminated. Examples of this type of data item include record codes, transaction codes, file identification items and update codes. Approximately three hundred of these types of data items were eliminated from the original set of eleven hundred.

(2) Systems analysts who were thoroughly familiar with the meanings and usage of data items were asked to describe the relationships between data items.

(3) The items were viewed as one relation of eight hundred domains using Codd's definition. Then the normalizing steps were followed, the complex relation was, in effect, broken down into thirty-nine simple relations written in third normal form. Each simple relation was given a name based on the real world entity which its data elements described. For example, the relation containing the data items which describe employees was called the employee entity. Each of the simple relations contains data items which describe attributes of an entity which is either: Physical, (for example, student, building), administrative, (for example, registration, account), or organizational, (for example, parent department). The thirty-nine entities yielded by the normalization process are as follows:

1. Advanced Standing Academic Record
2. Account
3. Accounting Balance
4. Alumni Membership
5. Applicant test Scores
6. Application
7. Application Control
8. Appointment
9. Authorization
10. Building
11. Business Address
12. Campus Address
13. Class Schedule
14. Course
15. Course Section
16. Courses Registered For
17. Current Month Accounting Transaction
18. Deduction History
19. Deductions
20. Degree Awarded
21. Donor Entity
22. Employee
23. Employee Class
24. Employee Tax Status
25. Expense Class
26. Home Address
27. Insurance and Retirement
28. Non-College Academic Record
29. Parent Department
30. Position Control
31. Promotional History
32. Registration
33. Room
34. Salary History
35. Skill and Education
36. Student/Applicant
37. Student Financial Data
38. Student's Family Data
39. Year-to-date Payment

Further information on the normalization process may be found in the texts already mentioned by Codd, Martin and Date.

Entity Grouping

The subjects for the entity grouping survey were selected through the process described below. Although the full survey instrument is not included here,

the instructions which subjects were given are shown below. The survey yielded a number of entity groupings which the subjects believed would be useful to them.

Seventy persons were identified by the Vice Presidents of the University to be interviewed in a long-range planning effort being carried out by the data processing department concurrently with this study. During the interviewing process, several interviewees indicated that they do not currently use data produced by the data processing department, nor did they see any future use for such institution-wide data. An example of this kind of situation is the Director of the University Press who essentially is running a small business which he manages primarily from data generated within his own department. The original seventy interviewees were reduced to fifty-four subjects for this survey. The group was made up of: all vice presidents, provosts, vice-provosts and deans; selected persons with primarily planning responsibilities and certain administrative staff persons.

The vice presidents originally selected this group of people as being representatives of their respective areas who ought to have input to the data processing department's long-range plan. It was therefore believed that they also would have the greatest needs for data from a future administrative database and therefore ought to have input to its design through this survey.

The survey was conducted by mailing to each subject a packet containing a deck of thirty-nine cards containing descriptions of each entity, survey instructions and response forms to be returned to the researcher. The instructions were as follows:

Step 1 Read all the steps of the process.

Step 2 Take a few minutes to identify the major decision areas in your administration in which data from the University's

database, either in summary or detail form, would be helpful.

Step 3 Describe each decision area on a separate decision form. Five forms have been provided for this purpose. You may select as many decision areas as you wish up to five.

Step 4 Indicate the number of times each year that you would need updated reports from the computer for this decision area in the space called "frequency of access".

Step 5 Look at each entity card and decide if the attributes of that entity would be useful data for the first decision area.

Step 6 For each entity which would be helpful for the first decision area, write the entity number on the decision form in the spaces called "Applicable Entity Numbers."

Step 7 Repeat steps 4 through 6 for each decision area.

Step 8 Return the decision forms.

Analysis of Administrator Grouping

A number of analysis techniques were tested including: observation of matrices showing the number of times each entity was grouped with each other entity; factor analysis of the groupings; and several cluster analysis techniques.

The matrices proved to be too large and complex to permit useful analysis by observation. Factor analysis of the groupings showed no useful clusters of entities and was therefore ruled out as an analysis method. Before discussing the results from cluster analysis methods tested, a brief discussion of cluster analysis may be useful.

Blashfield suggests that there are probably one hundred different clustering techniques available today, each described in terms unique to the field in which the method originated (Blashfield, 1976). Each clustering technique has characteristics which may cause it be be more or less useful for clustering a given set of data. The impact of the characteristics of a given method on a given set of data

cannot always be predicted so that results of a given cluster analysis must be tested for their validity. Further, Blashfield recommends that rather than select a particular clustering method the researcher may wish to try several methods and compare the results before selecting one method. In this way, the researcher can analyze the differences between results to decide which method has characteristics which will be most useful for the analysis.

Blashfield describes four clustering methods in his article each of which was used in the analysis of the administrator groupings. The differences between the methods are in the way that each links entities into groups or clusters. All four methods are hierarchical clustering methods which give an output format suited to the analysis required here. The four methods are:

Single linkage	Each entity is linked into the cluster which it is most like, or in this case, used with most frequency.
Complete linkage	Each entity links to the cluster containing all entities which are more similar to it than are all the entities of any other cluster, where similarity is measured by the frequency with which entities were used together.
Average linkage	Each entity links with the cluster whose average value is most similar to it.
Minimum Variance or Ward Method	This method clusters entities in such a way that the variance between entities within a cluster in minimized.

Anderberg has written a series of programs which will perform hierarchical cluster analysis by seven different methods including those mentioned above (Andberg, 1973). In this study all four methods were used, the results of which will be described in the Results section. The complete linkage and average

linkage methods appear to yield the most reasonable solution based on comparison of these solutions to current data file structures.

Results

The data structures as currently used may be approximately divided into seven major groups: employee, facility, address, course, student, financial and development. These structures were compared to the cluster analysis to determine the reliability of the results as well as relationships which appear in the results but are not in current structures. Although a number of different clusterings were computed and each provided additional insight to the probable future relational requirements of top administrators, only one clustering result is presented for discussion here (see Baltes for detailed discussion of the analysis and findings of the study).

Figure I below shows one hierarchical clustering of the entity groupings. The strength of a relationship between two entities or one entity and a cluster may be measured by the point at which they join in the hierarchy where entities with strong relationships (used together with high frequency) will connect on the left hand side while those with weaker relationships will connect further towards the right. For example, administrators grouped employee tax status with year-to-date payment data very often while employee tax status was almost never grouped with registration data. Ultimately, by design, this analysis method will always link every entity in the groupings.

Examination of the cluster results shows that only five separate groupings (divided by the dash line) can be identified compared with the seven in current structures. The top cluster compares roughly to the employee file. The second

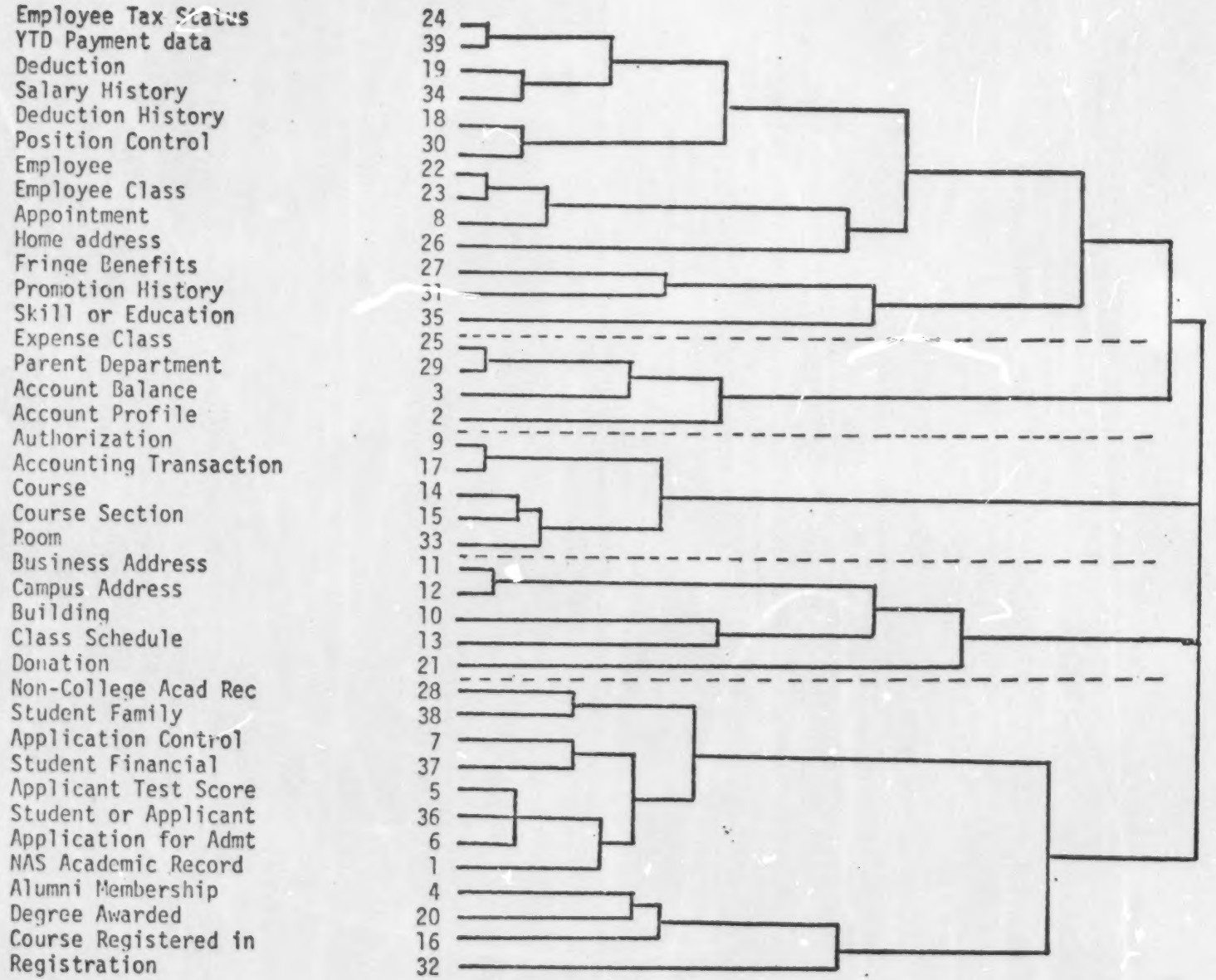
Employee Tax Status 24
YTD Payment data 39
Deduction 19
Salary History 34
Deduction History 18
Position Control 30
Employee 22
Employee Class 23
Appointment 8
Home address 26
Fringe Benefits 27
Promotion History 31
Skill or Education 35
Expense Class 25
Parent Department 29
Account Balance 3
Account Profile 2
Authorization 9
Accounting Transaction 17
Course 14
Course Section 15
Room 33
Business Address 11
Campus Address 12
Building 10
Class Schedule 13
Donation 21
Non-College Acad Rec 28
Student Family 38
Application Control 7
Student Financial 37
Applicant Test Score 5
Student or Applicant 36
Application for Admt 6
NAS Academic Record 1
Alumni Membership 4
Degree Awarded 20
Course Registered in 16
Registration 32

FIGURE I

cluster is approximately comparable to the current financial file. The third cluster compares to the current course file. The fourth cluster combines the current address and development files and the fifth relates to the current student file.

The summary statements drawn from this analysis are:

(1) Top administrators view alumni data as part of the student file rather than the development file as current structures carry the data.

(2) Top administration is not concerned about facilities data except as it relates to the course schedule and campus address.

(3) Authorizations are an accounting entity related to building construction and maintenance. Although current structures carry authorization data in the accounting system, top administration appears to see authorization as strongly related to room. (Current data processing systems can relate authorization data with rooms only with great difficulty and expense).

Although the above comments relate to one clustering, a number of them were done. As each of the clusterings are examined, additional relationships not supported by current systems were uncovered - relationships which future databases must support if they are to be responsive to management's needs.

Conclusions

The methodology was proven to be a very useful tool for the database administrator for anticipating future relational requirements of the databases he designs. Some difficulties were encountered in getting responses and the ability of some subjects to understand the instructions. An interview format might have resolved some of these problems.

The implications of this study for database designers must be viewed from the perspective of a total systems design scenario. The database designer must

know a good deal more about the data resource than was necessary to design a conventional single application file. He must understand: the logical views which the database must be able to present for all users; security and privacy constraints for all entities in the database; pecularities of the database system which will be used and their impacts on the physical implementation of the database; the value of the data as a resource of the institution which will have implications on what the institution will demand in terms of efficiency of use and frequency of use; and many other aspects which I have not mentioned. This methodology could be effective not only in building a relational model of the database reflecting top administration's view ofthe data but for finding all logical views that users may require of the database. If prospective users were asked to group entities not by decision area but by frequency of use or level of security or privacy constraint and so on, the database administrator could cluster each set of responses to get different perspectives of the data, all of which will need to be taken into consideration in the final implementation of the database.

REFERENCES

1. Anderberg, M. R. Cluster Analysis for Applications. New York: Academic Press, 1973.

2. Baltes, K. G. Data Needs Assessment for Higher Education Administration: A Methodology. Unpublished Ph.D. dissertation, University of Minnesota, 1977.

3. Blashfield, R. K. Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods. Psychological Bulletin, 1976, 83 (3), pp. 377-388.

4. Codd, E. F. A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, June 6, 1970, 13, pp. 377-787.

5. Date, C. J. An Introduction to Data Base Systems. Reading, Massachusetts: Wesley Publishing Co., Inc., 1975.

6. Martin, J. Computer Data Base Organization. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1975.